Claims

1. An audio processing system for providing synchronized display of recognized text from an original audio file containing speech spoken by a user and playback of the original audio file, said system comprising:

(a) a speech recognition module for generating recognized text pieces and associated audio pieces from the original audio file;

(b) a silence insertion module for aggregating the audio pieces into an aggregated audio file;

(c) a silence detection module for converting the original audio file and the aggregated audio file into a silence detected original audio file and a silence detected aggregated audio file, wherein silent and non-silent groups are identified using a threshold volume;

(d) said silence insertion module further being adapted to:

(i) compare the silence detected original audio file with the silence detected aggregated audio file and determine the differences in position of the non-silence group within the respective files;

(ii) insert silence within the audio pieces according to the differences in position determined in (i) to create silence inserted audio pieces, such that aggregation of the silence inserted audio pieces results in an aggregated silence inserted audio pieces file that substantially corresponds to the original audio file; and

(iii) utilize the characteristics of the silence inserted audio pieces and the associated recognized text pieces to synchronize the display of the recognized text pieces from the original audio file and the playback of the associated audio pieces from the original audio file.

2. The system of claim 1, wherein said silence detection module further utilizes an adaptive sliding average window register that maintains the average volume of a number of preceding non-silence groups to determine whether a following non-silence group should be considered as non-silence or as silence.

3. The system of claim 1, wherein each of said non-silence groups have a height that corresponds to the average volume of the audio within said blocks.

4. The system of claim 1, wherein the threshold volume is set to reflect the recording environment for the original audio file.

5. The system of claim 1, wherein said silence insertion module is further adapted to:

(i) assign an original audio marker to a position in front of a non-silence group in the silence detected original audio file and an aggregated marker to a position in front of a non-silence group in the silence detected aggregated audio file;

(ii) determine if the respective non-silence groups match;

(iii) determine the difference in position between said original audio marker and said aggregated marker;

(iii) repeat (i).

6. The system of claim 1, wherein the characteristics of a silence inserted audio piece and an associated recognized text piece include at least one of the group consisting of: the position of the starting character of the text piece in the audio piece relative to the start of original audio file, the number of characters in the text piece, the duration of the audio piece, the confidence of the recognized text piece.

7. The system of claim 1, wherein the silence insertion module synchronizes the display of the recognized text pieces and the playback of the associated audio pieces by utilizing the timing characteristics of the silence inserted audio pieces.

8. The system of claim 7, wherein the silence insertion module synchronizes the display of the recognized text pieces and the playback of the associated audio pieces by highlighting a recognized text piece and at least one adjoining recognized text piece while playing back the audio pieces that correspond to said recognized text piece and at least one adjoining recognized text piece.

9. The system of claim 1, further comprising a terminal coupled to said speech recognition module, said silence detection module and said silence insertion module, said terminal being utilized to obtain said original audio file from the user.

10. The system of claim 1, further comprising a playback device coupled to said silence insertion module for receiving the aggregated silence inserted audio pieces file and for displaying the recognized text pieces and playing back the associated audio pieces in a synchronized manner.

11. An audio processing method for providing synchronized display of recognized text from an original audio file containing speech spoken by a user and playback of the original audio file, said method comprising:

(a) recognizing the spoken speech within the original audio file and generating recognized text pieces and associated audio pieces;

(b) aggregating the audio pieces into an aggregated audio file;

(c) applying silence detection to convert the original audio file and the aggregated audio file into a silence detected original audio file and a

silence detected aggregated audio file, wherein silent and non-silent groups are identified using a threshold volume;

(d) comparing the silence detected original audio file with the silence detected aggregated audio file and determining the differences in position of corresponding non-silence groups within the silence detected original audio file and the silence detected aggregated audio file;

(e) inserting silence within the audio pieces according to the differences in position of corresponding non-silence groups within the silence detected original audio file and the silence detected aggregated audio file to create silence inserted audio pieces, such that aggregation of the silence inserted audio pieces results in an aggregated silence inserted audio pieces file that substantially corresponds to the original audio file; and

(f) utilizing the characteristics of the silence inserted audio pieces and the associated recognized text pieces to synchronize the display of recognized text from an original audio file and playback of original audio file.

12. The method of claim 11, wherein step (c) includes the step of utilizing an adaptive sliding average window register to maintain the average volume of a number of preceding non-silence groups to determine whether a following non-silence group should be considered as non-silence or as silence.

13. The system of claim 11, wherein each of said non-silence groups have a height that corresponds to the average volume of the audio within said blocks.

14. The method of claim 11, wherein step (c) further includes determining the threshold volume that reflects the recording environment for the original audio file.

15. The method of claim 11, wherein step (d) further comprises:

(i) assigning an original audio marker to a position in front of a non-silence group in the silence detected original audio file and an aggregated marker to a position in front of a non-silence group in the silence detected aggregated audio file;

(ii) determining if the respective non-silence groups match;

(iii) determining the difference in position between said original audio marker and said aggregated marker;

(iii) repeat (i).

16. The method of claim 11, wherein the characteristics of a silence inserted audio piece and an associated recognized text piece of step (f) include at least one of the group consisting of: the position of the starting character of the text piece in the audio piece relative to the start of original audio file, the number of characters in the text piece, the duration of the audio piece, the confidence of the recognized text piece.

17. The method of claim 11, wherein step (f) further includes synchronizing the display of the recognized text pieces and the playback of the associated audio pieces by utilizing the timing characteristics of the silence inserted audio pieces.

18. The method of claim 11, wherein step (f) includes synchronizing the display of the recognized text pieces and the playback of the associated audio pieces by highlighting a recognized text piece and at least one adjoining recognized text piece while playing back the audio pieces that correspond to said recognized text piece and at least one adjoining recognized text piece.

19. The method of claim 11, wherein said original audio file is obtained from the user using a terminal.

20. The method of claim 11, wherein the recognized text pieces are displayed and the associated audio pieces are played back in a synchronized manner with the recognized text pieces using a playback device.